DOCUMENT RESUME

ED 262 056 TM 850 531

AUTHOR Crovo, Mary L.; Phillips, Gary W.
TITLE Detection of Bias Test Items Using Item Response Theory Models.
PUB DATE Apr 83
NOTE 22p.; Paper presented at the Annual Meeting of the American Educational Research Association (67th, Montreal, Quebec, April 11-15, 1983). Appendices contain faint, broken type.
PUB TYPE Speeches/Conference Papers (150) -- Reports - Descriptive (141) -- Tests/Evaluation Instruments (160)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Advisory Committees; Criterion Referenced Tests; *Culture Fair Tests; *Item Analysis; *Latent Trait Theory; Minimum Competency Testing; Racial Bias; Secondary Education; Sex Bias; State Programs; Statistical Analysis; Student Certification; *Test Bias; *Test Construction; Testing Programs; Test Items
IDENTIFIERS Hambleton (Ron K); *Maryland Functional Mathematics Test; *Maryland Functional Reading Test

ABSTRACT

This paper presents the dual approach to item bias detection employed in the Maryland Functional Testing Program (MFTP). Using instructional objectives mandated by the Maryland State Board of Education, the MFTP develops two levels of the Maryland Functional Reading Test (MFRT) and the Maryland Functional Mathematics Tests (MFMT). These multiple-choice, domain-referenced instruments are used for high school graduation certification (grade 9 version) and for diagnostic purposes (grade 7 version). The item bias detection process involves a 15-member Bias Committee which reviews items prior to field testing and then, one year later, items are reviewed subsequent to field testing accompanied by statistical bias data which uses Hambleton's graphical analysis and replication technique. As indicated by statistical and graphical results, no items on the Level I or Level II MFRT appeared racially or sexually biased. Item review form and graphs showing race bias analysis conclude the document. (BS)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED262056

DETECTION OF BIASED TEST ITEMS USING ITEM RESPONSE THEORY MODELS

BY

MARY L. CROVO

AND

GARY W. PHILLIPS

MARYLAND STATE DEPARTMENT OF EDUCATION

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

B. W. Phillips

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

A symposium presentation at the annual meeting of the American Educational Research Association, Montreal, April 11-15, 1983.

TM 850 531

The development of measurement instruments which accurately assess examinee ability, regardless of group affiliation, is of serious concern to the testing specialist. This concern is particularly critical in the area of competency-based testing, where graduation certification is contingent upon student test performance.

Various techniques for detecting biased test items have appeared in the literature over the past decade. However, no single methodology has emerged as the "best" bias detection procedure.

The purpose of this paper is to present the dual approach to item bias detection employed in the Maryland Functional Testing Program (MFTP). This approach utilizes both empirical and judgmental techniques for identifying biased test items.

The primary goal of the MFTP is to create and administer instructionally relevant and sensitive measurement instruments. The Maryland Functional Reading Test (MFRT) and the Maryland Functional Mathematics Test (MFMT) are multiple-choice, domain-referenced instruments used to certify high school graduation beginning in grade nine. A lower level version of each test is administered in grade seven solely for diagnostic purposes. State Board mandated and adopted instructional objectives serve as the content base for these two areas of assessment in the MFTP. These objectives are now an integral part

of the instructional programs in each local educational agency in Maryland. The actual test items are written based upon detailed test specifications developed by Maryland content and measurement people.

The test development timeline for each of these assessment instruments follows a rigorous 28-month calendar, including specifications review, item generation, content reviews, field testing, Rasch calibrating and equating, then further content reviews, until the items are finally chosen for an operational test form. Insuring that no biased items appear on an operational test form involves a multi-stage statistical and judgmental review process.

Bias Committee

An integral part of the Maryland bias detection process involves a 15-member Bias Committee. The composition of this Committee assures that the areas of race, sex, handicapping condition, cultural and geographical bias are represented. Persons familiar with bias in Maryland within these areas, from both the public and private sector, meet for a one-week workshop. The charge to this Committee is to review items prior to field testing and then, one year later, items are reviewed subsequent to field testing accompanied by statistical bias data.

Prior to the actual item bias review, the Committee is provided with a background on the MFTP developmental process. By establishing the context in which the tests are constructed and the purposes which they must serve, the Committee can better focus on its task and its

importance in the process.

The first screening of items by the Committee prior to field testing is solely a judgmental bias review. An item bias review form developed by Hambleton (1980) is used to detect items which may reflect sexual, cultural, racial, regional and/or ethnic content bias (see Figure 1). Approximately 200 reading and 200 mathematics items are screened for bias by the Committee with each item being screened by at least two persons. Items flagged as biased are discussed by the Committee during the week-long meeting. Then the items agreed upon as biased are returned for content editing or are deleted from the item bank completely, depending on the seriousness of the biased content. The unbiased items are placed in a bank from which field test items are randomly drawn. These field test items are embedded in alternate forms which are administered to students in the fall.

The following spring, the Bias Committee is reconvened to review those field tested items which were flagged as biased by statistical procedures. (The statistical bias procedures used to flag the items will be outlined in the next section of this paper).

This session involves both judgmental and statistical input. Those field test items flagged as biased are examined by the Committee to determine why the performance data indicated a differential performance for a certain subgroup. If possible, these items are then returned to the item banks with content editing to eliminate the bias.

The flowchart in Figure 2 depicts the major steps in the MFTP

development surrounding the Bias Committee. According to this two-stage screening process, an item must first pass a judgmental review prior to its appearance in a field test position. Subsequent to field testing, items flagged as biased are subjected to a judgmental and empirical review. Items which are agreed upon as biased are salvaged, whenever possible, and returned to the review cycle. Thus, no item is ever placed on an operational test form without first passing this dual screening process.

Statistical Procedures

A wide-ranging menu of statistical item bias methods was reviewed prior to the selection of procedures for diagnosing bias on the Maryland Functional Reading and Mathematics Tests (Berk, 1982). In the past, a variety of methods were used to assess these instruments for sexual and racial bias. The major techniques most often employed were those recommended by Scheuneman (1979). However, it became immediately apparent that all of the common methods employed had two major shortcomings, in that they were unable to control for: 1) the confounding effects of differential ability distributions, and 2) sample sizes in the compared groups.

Hambleton (1982) suggested a simple solution to this problem by employing a graphical analysis and replication methodology to detect item bias. This technique matches all groups on ability and sample size prior to estimating the item parameters. The Maryland Functional Testing Program now uses this procedure along with methods recommended by Wright, Mead and Draba (1976) to assess item bias.

Before presenting an outline of Hambleton's graphical analysis technique, a brief discussion of the psychometric properties of the Maryland Functional Reading and Mathematics Tests is in order. The one parameter Rasch model is used to calibrate the test items and to equate forms from year to year. Once field test Rasch item information is available, item selection proceeds based on content and statistical criteria. Items with Rasch between t-fit statistics of 4.00 or greater are deleted from the bank. Likewise, items are removed which are not judged to be content valid according to stringent item-test specifications congruence criteria. The item bank is then recalibrated, and test construction continues. Thus, the bank contains only items which fit the one parameter Rasch model and are absolutely congruent with the test specifications.

Hambleton's graphical analysis and replication technique was chosen because of its inherent simplicity in terms of computation and interpretation. It also does not fall prey to common criticisms regarding sample size, which plague methods employing significance tests. Both the reading and mathematics items are investigated for potential sexual and racial bias in this manner.

Hambleton's procedure may be illustrated by the following steps used to assess sexual item bias:

1. Male and female groups are each randomly divided into two subgroups. This results in groups A and B for males, and groups A and B for females. This procedure provides a random replication for each comparison.

2. Group A males and females are identically matched on the total test raw score. Likewise, group B males and females are identically matched. This procedure insures that male and female comparisons are being made with identical sample sizes and ability distributions.
3. All four groups (male A, male B, female A and female B) are separately calibrated using the Rasch model.
4. Item difficulty estimates are compared in all four groups. Since the same people take the same items in all four groups, the item calibrations are automatically equated.

The above procedure provides three important pieces of information: 1) how much sex bias is present in each item (using group A males versus group A females; 2) how much of the detected sex bias shows up on a random replication (using group B males versus group B females); and 3) how much of this detected bias is really random error (using group A males versus group B males pooled with group A females and group B females).

Statistical Results and Discussion

Statistical results were available for the detection of item bias on the Maryland Functional Reading Test (MFRT) for both Level I and Level II data. Table 1 presents the sample sizes used in the analyses for racial and sexual bias of the MFRT items. Due to the matching procedure, sample sizes are equal for groups within Level I and Level II for race and sex, respectively. Table 2 lists the correlation coefficients and standard errors of estimate obtained from the Level I and Level II race and sex bias plots. To illustrate these results

graphically, Figures 3 - 6 depict the four plots for the Level I sexual bias analyses: 1) Group A Blacks verses Group A Whites; 2) Group B Blacks versus Group B Whites; 3) Group A Blacks versus Group B Blacks; and 4) Group A Whites versus Group B Whites.

As indicated by the statistical and graphical results, no items on the Level I or Level II Maryland Functional Reading Test appeared racially or sexually biased. All correlations and corresponding plots of item difficulty estimates show nearly perfect linear relationships between the matched groups of examinees, in both sex and race analyses. Since the items were selected on the basis of fit to the one parameter Rasch model, these results are not surprising. For an item to be flagged as potentially biased, it would have appeared as an outlier in the first plot of Group A males verses Group A females, for example. The same item would also have been an outlier in the replication plot of group B males versus Group B females.

Conclusions

Hambleton's graphical analysis and replication technique provides a conceptually simple, yet elegant approach to the detection of biased test items. By coupling the judgmental reviews of Maryland's Bias Committee with a sound empirical methodology, the detection of biased items in the Maryland Functional Testing Program can proceed with confidence. Further research in this technique is needed to determine the potentially biased outlier items in an empirical manner. Applications of this graphical analysis technique would be invaluable to other measurement areas in addition to item bias detection.

TABLE 1

Sample Sizes Used in Race and Sex Bias Analysis of the MFRT

	Race		Sex	
	Black	White	Male	Female
Level I				
Group A	1576	1576	12014	12014
Group B	1582	1582	11888	11888
Level II				
Group A	2186	2186	13910	12910
Group B	2194	2194	13680	13680

TABLE 2

Correlation Coefficients and Standard Errors of Estimate for Racial and Sexual Bias Plots

Racial Bias	Level I	Level II
Group A Blacks vs. Group A Whites	.98[a]	.98
	.20[b]	.22
Group B Blacks vs Group B Whites	.98	.98
	.20	.26
Group A Blacks vs Group B Blacks	.99	.99
	.12	.13
Group A Whites vs Group B Whites	.99	.99
	.12	.12

Sexual Bias	Level I	Level II
Group A Males vs Group A Females	.99	.99
	.19	.18
Group B Males vs Group B Females	.99	.99
	.18	.19
Group A Males vs Group B Males	1.00	1.00
	.05	.06
Group A Females vs Group B Females	1.00	1.00
	.06	.06

a: Correlation coefficient
b: Standard error of estimate

Figure 1

Item Review to Detect Bias[1]

Reviewer: _____ Date: _____ Objective: _____

This review form has been designed to assist in the identification of test items which may reflect sexual, cultural, racial, regional and/or ethnic content bias and stereotyping. Place your name, the date, and the objective number at the top of the review form.

In the spaces at the right under the heading "Test Item Number" provide the numbers of the test items you will review. When the number of test items for review exceeds ten, use a second review form.

Next, read each test item and answer the six questions below. Use "✓" for YES, "X" for NO, "?" for UNSURE, and "NA" for NOT APPLICABLE. When your rating is negative ("X" or "?"), indicate what you think is the problem in the test item and suggest a revision on your copy of the test item.

When you have completed your review task, staple the test items to your review form(s) and return.

	Test Item Number									
1. Is the item free of offensive sexual, cultural, racial, regional and/or ethnic content?										
2. Is the item free of sexual, cultural, racial, and/or ethnic stereotyping?										
3. Is the item free of language which could be offensive to a segment of the examinee population?										
4. Is the item free from descriptions which could be offensive to a segment of the examinee population?										
5. Will the activities described in the item be equally familiar (or equally unfamiliar) to all examinees?										
6. Will the words in the item have a common meaning for all examinees?										

[1]Prepared by Ronald K. Hambleton from the University of Massachusetts, Amherst. Comments on the soundness of this draft instrument would be greatly appreciated.

BEST COPY AVAILABLE

Figure 2

BIAS COMMITTEE REVIEW SEQUENCE

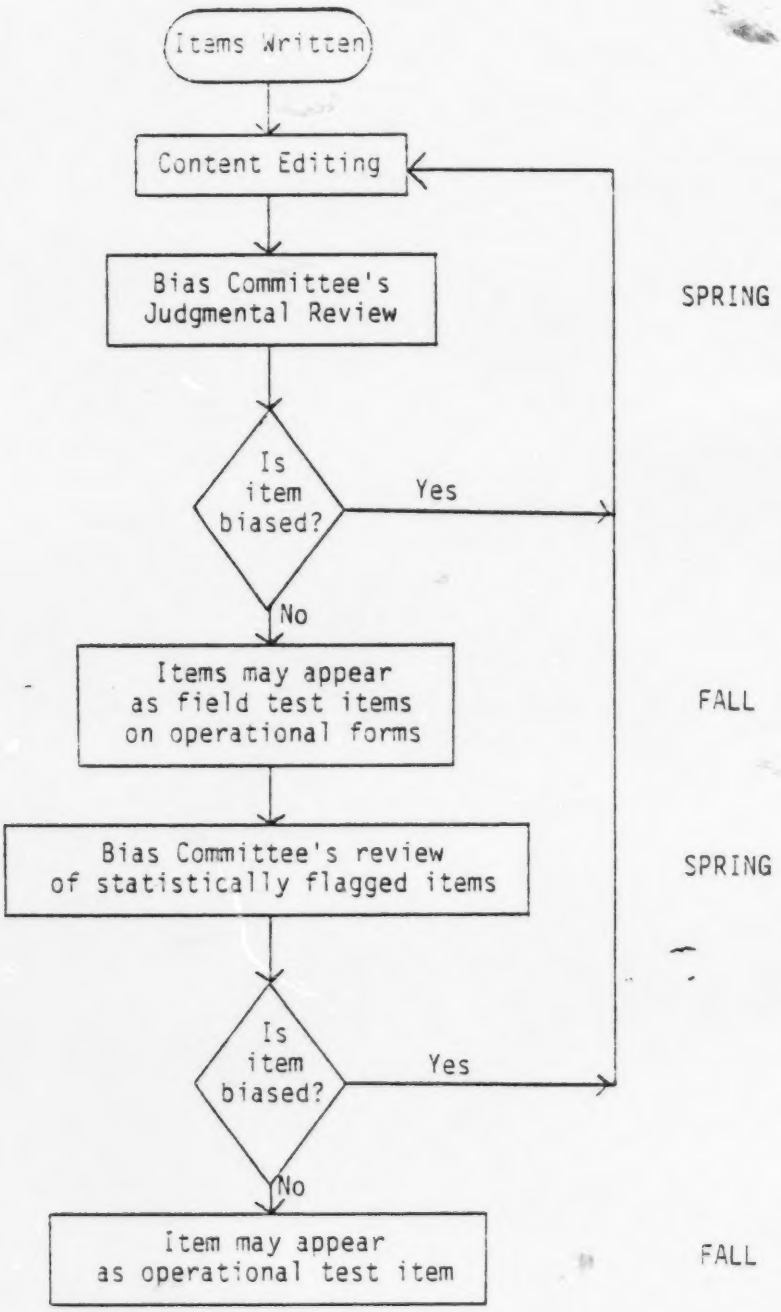
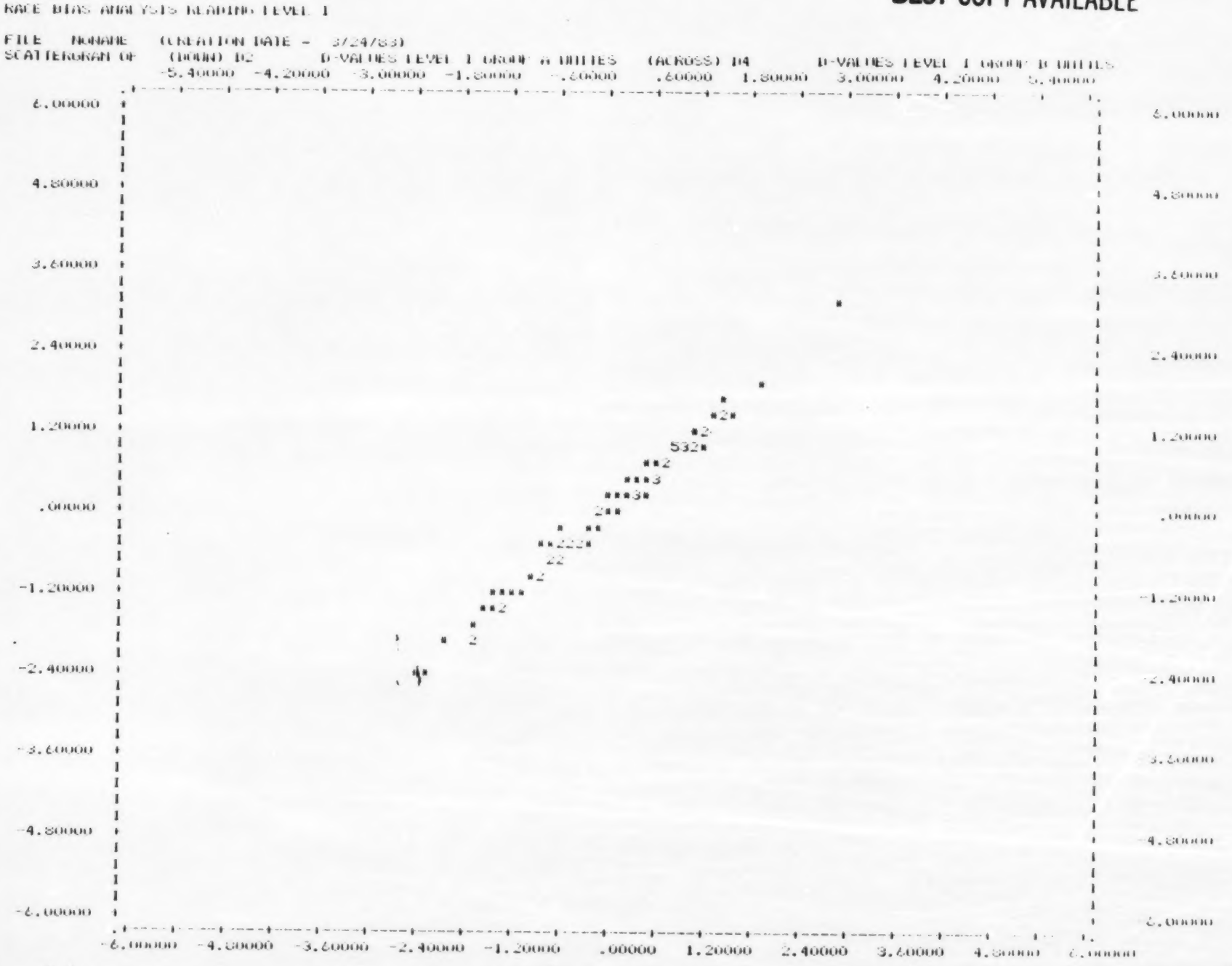


FIGURE 3

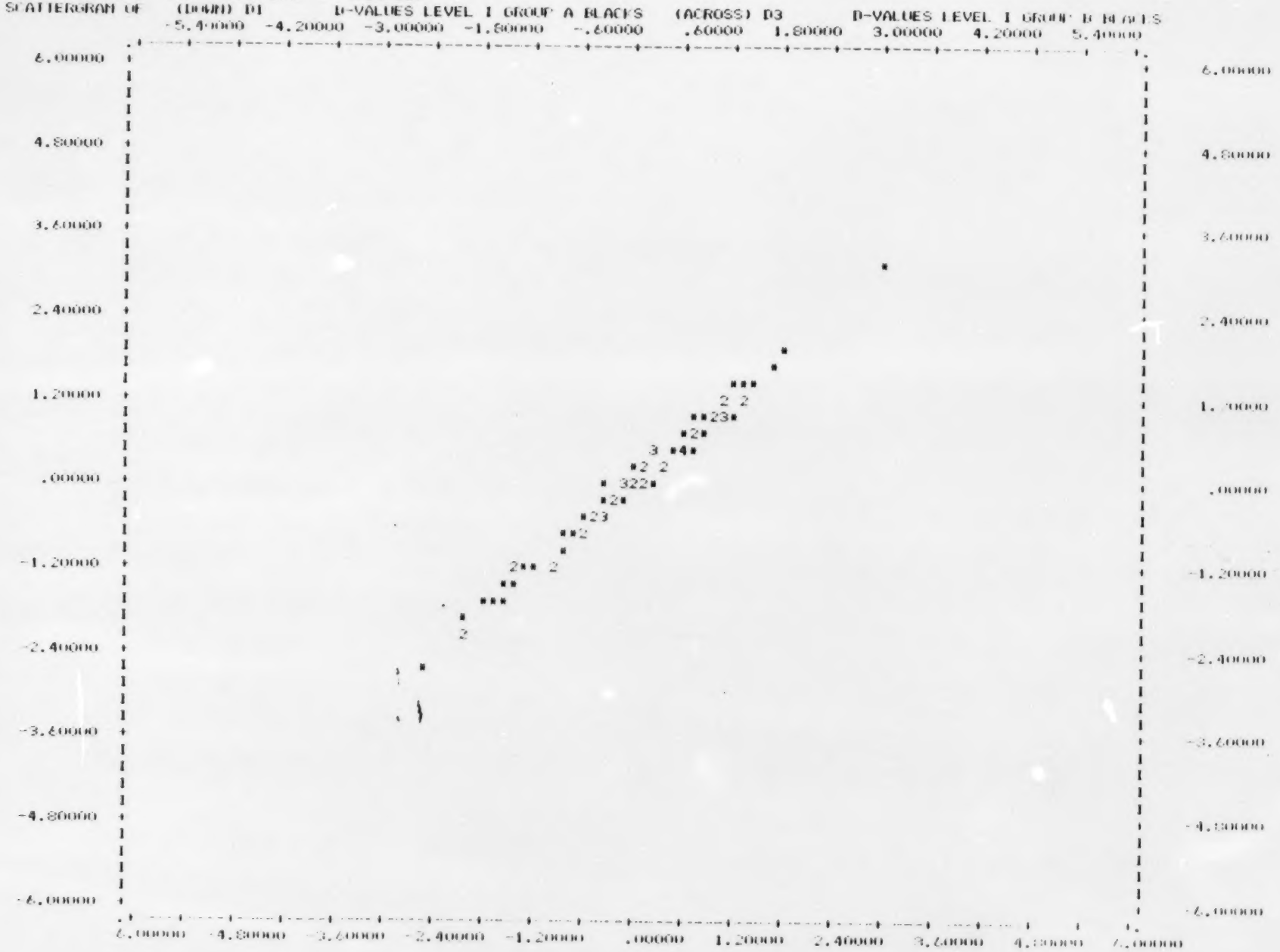
BEST COPY AVAILABLE



RACE BIAS ANALYSIS READING LEVEL I

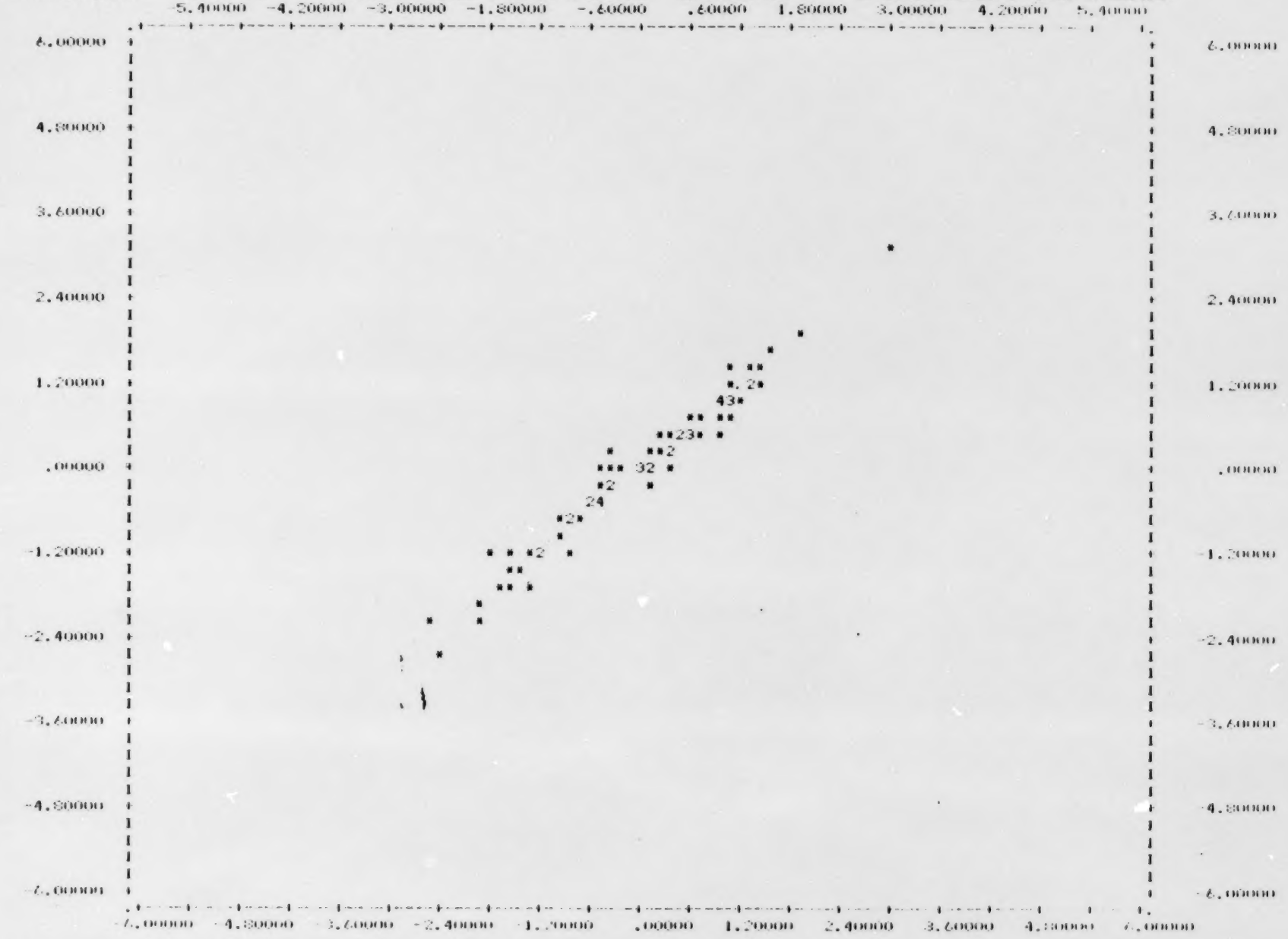
FIGURE 4

BEST COPY AVAILABLE



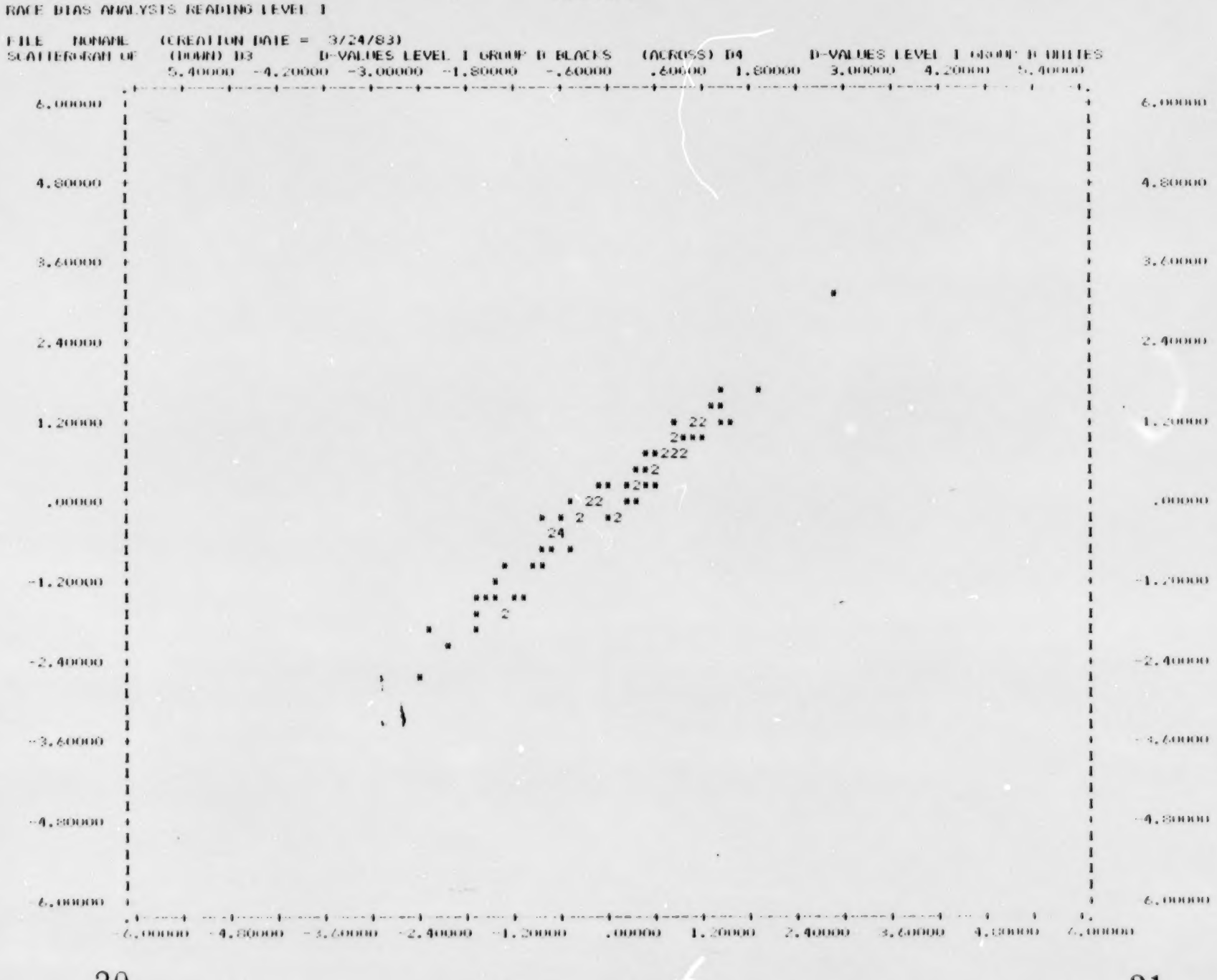
RACE BIAS ANALYSIS READING LEVEL I

FIGURE 5



BEST COPY AVAILABLE

FIGURE 6


RACE BIAS ANALYSIS READING LEVEL 1
FILE NONAME (CREATION DATE = 3/24/83)
SCATTERGRAM OF (DOWN) D3 D-VALUES LEVEL 1 GROUP B BLACKS (ACROSS) D4 D-VALUES LEVEL 1 GROUP B WHITES
5.40000 -4.20000 -3.00000 -1.80000 -.60000 .60000 1.80000 3.00000 4.20000 5.40000
6.00000
4.80000
3.60000
2.40000
1.20000
.00000
-1.20000
-2.40000
-3.60000
-4.80000
-6.00000
-6.00000 -4.80000 -3.60000 -2.40000 -1.20000 .00000 1.20000 2.40000 3.60000 4.80000 6.00000

REFERENCES

Berk, R.A. Handbook of methods for detecting test bias. Baltimore: The Johns Hopkins University Press, 1982.

Hambleton, R.K. Item review to detect bias. Prepared for the University of Massachusetts, Amherst, 1981.

Hambleton, R.K. Some goodness of fit investigations for item response models. To be published in Hambleton, R. K. (Ed.), Applications of item response theory. Vancouver, British Columbia: Educational Research Institute of British Columbia, 1982.

Rudner, L. The use of latent trait theory for investigating cross-cultural test item behavior. A paper presented at the annual meeting of the Eastern Educational Research Association, 1981.

Scheuneman, J. A method of assessing bias in test items. Journal of Educational Measurement, 1979, 16,3,143-152.

Wright, B.D., Mead, R.M., and Draba, R. Detecting and correcting test item bias with a logistic response model (Research Memorandum No. 22) Chicago: Statistical Laboratory, Department of Education, University of Chicago, 1976.

BEST COPY AVAILABLE